



Early Journal Content on JSTOR, Free to Anyone in the World

This article is one of nearly 500,000 scholarly works digitized and made freely available to everyone in the world by JSTOR.

Known as the Early Journal Content, this set of works include research articles, news, letters, and other writings published in more than 200 of the oldest leading academic journals. The works date from the mid-seventeenth to the early twentieth centuries.

We encourage people to read and share the Early Journal Content openly and to tell others that this resource exists. People may post this content online or redistribute in any way for non-commercial purposes.

Read more about Early Journal Content at <http://about.jstor.org/participate-jstor/individuals/early-journal-content>.

JSTOR is a digital library of academic journals, books, and primary source objects. JSTOR helps people discover, use, and build upon a wide range of content through a powerful research and teaching platform, and preserves this content for future generations. JSTOR is part of ITHAKA, a not-for-profit organization that also includes Ithaka S+R and Portico. For more information about JSTOR, please contact support@jstor.org.

THE MATHEMATICS TEACHER

EDITED BY
W. H. METZLER

ASSOCIATED WITH
EUGENE R. SMITH HARRY D. GAYLORD
MAURICE J. BABB WILLIAM E. BRECKENRIDGE

VOLUME VI	MARCH, 1914	NUMBER 3
-----------	-------------	----------

AN EXPERIMENT IN GRADING PROBLEMS IN ALGEBRA.

BY EDWARD L. THORNDIKE.

THE DIFFICULTY OF CERTAIN PROBLEMS IN ALGEBRA AS JUDGED BY THE CONSENSUS OF TWO HUNDRED TEACHERS OF MATHEMATICS.

Two hundred teachers of mathematics, chiefly members of the New York Section of the Association of Teachers of Mathematics of the Middle States and Maryland, ranked the twenty-five problems printed below for difficulty, "difficulty" being defined as in the instructions appended. The variations in the individual opinions were very great, being as shown in Table I. It is an interesting exercise to examine this table, and imagine, as well as one can, the points of view from which these varying estimates were each plausible—to divine, for example, why Problem *T* was rated all the way from easiest to hardest of the twenty-five. How much of the variation was due to tenable points of view and how much was due to errors of judgment cannot, of course, be told until the problem in question has been tested with respect to the percentage of pupils able to solve it in the time allowed.

In spite of the great variation, the consensus of the two hundred teachers gives a fairly clear order of difficulty for the problems, as shown at the bottom of Table I. *D* is easiest; *K*

TABLE I.

THE RELATIVE POSITIONS ASSIGNED TO EACH OF 25 PROBLEMS IN ALGEBRA BY 200 TEACHERS OF MATHEMATICS.

The entries in the body of the table give the frequencies of each problem in each position. The first line below the body of the table gives the position or point between two positions which most nearly divides the ratings into halves. The next three lines give the results of the division. The last line gives the most probably true order of difficulty by the consensus.

Position	+D	K	A	L	O	S	T	P	U	G	B	C	Q	H	R	E	M	I	X	Y	J	N	V	F	W
1	146	32	7	8			3			1									1						1
2	43	121	19	10	2		3		1																2
3	8	21	75	25	52	7	3		3		4								1						3
4	1	17	27	65	26	41	8		5	3	3								1						4
5	1	4	26	26	63	19	21	12	11	7	7														5
6	1	3	16	23	26	61	16	11	14	8	12	1							2						6
7		1	16	11	17	31	24	32	15	27	10	4													7
8		1	5	17	10	15	22	45	28	18	18	7							1						8
9			5	8	2	10	14	43	33	24	16	20							3						9
10			2	4		3	16	14	16	21	16	47							1						10
11			1			1	11	8	15	12	14	33							1						11
12						2	1	12	6	20	9	24							5						12
13							3	7	1	11	7	13							3						13
14							5				9	19							2						14
15							4	1	2	8	11	7							3						15
16							5			9	6	6							6						16
17							6		1	3	13	1							11						17
18							2			5	12	2							16						18
19							2		1	6	2								31						19
20							1			1	2								28						20
21							1			1									32						21
22							1			2									30						22
23							1			1									14						23
24							1			2									12						24
25							1			2									8						25
Approximate Median Position	1	2	3-4	4-5	5	6	8-9	8-9	9-10	10	10-11	11-12	13-14	13-14	15-16	15-16	16-17	19-20	19-20	19-20	20	21-22	22	23-24	24-25
Data on median position	32	101	108	80	67	100	102	107	91	97	108	109	101	103	98	106	101	101	101	97	89	105	82	107	100
Order of difficulty by consensus	47	121	99	92	63	61	100	98	93	20	103	92	91	99	97	102	94	99	99	103	24	95	43	93	100
	1	2	3	4	5	6	7 and 8	9	10	11	12	13	14	15	16	17	18, 19 and 20	18, 19 and 20	18, 19 and 20	21	22	23	24	25	

is next easiest; A is next; and so on (T and P being closely alike and I , X and Y being closely alike). These position-values are, however, of comparatively little use. The significance of any one of them depends upon consideration of the entire series, and is, as it stands, only a means of relative position, not of the amount of "difficulty" or of anything else.

LIST OF PROBLEMS.

To be ranked in order of difficulty. See sheet of instructions.

A. If $x + 3a = 5a$, what does x equal?

B. The circumference of a circle $= 2\pi r$. $\pi = 3\frac{1}{7}$. r = the length of the radius of the circle in question. If the diameter of a bicycle wheel is 28 inches, how many inches is the circumference?

C. If $\frac{6x + 7}{5} - \frac{2x - 1}{10} = 4\frac{1}{2}$, what does x equal?

D. If $a = 4$ and $b = 2$, what does $a + b$ equal?

E. If $2 + \frac{\frac{x}{a} - 1}{\frac{2}{a}} = 0$, what does x equal?

F. A cube containing eight cubic inches was plated with copper. The difference in the weights of the cube before and after the plating was 0.139 lbs. 1 cubic inch of copper weighs 0.315 lbs. Form an equation from which the approximate thickness of the copper plating could be calculated. State whether the approximate estimated thickness by your equation would be less or more than the exact thickness.

G. If $a = 6$ and $b = 3$, what does $\sqrt{a} \sqrt{2b}$ equal?

H. If $\frac{1}{a} - \frac{1}{x} = \frac{1}{x} - \frac{1}{b}$, what does x equal?

I. A man has a hours to spend riding with a friend. How far can they ride together, going out at the rate of b miles an hour, and just covering the return trip at the rate of c miles an hour?

J. If $\frac{a+b}{b+c} = \frac{c+d}{d+a}$, prove that $a=c$ or that $a+b+c+d=0$.

K. If $a=4$ and $b=0$, what does $a+b$ equal?

L. If $3x+4=2x+8$, what does x equal?

M. If $\frac{x+a}{x-a} - \frac{x-a}{x+a} - \frac{x^2}{a^2-x^2} = 1$, what does x equal?

N. There are two thermometers or scales to measure temperature. The Fahrenheit scale (F.) is the one we commonly use. The other is called the Centigrade scale (C.). A temperature of 32 degrees on the F. scale $= 0$ degrees on the C. scale. 33.8 degrees on the F. scale $= 1$ degree on the C. scale. 35.6 degrees on F. $= 2$ degrees on C. 50 degrees on F. $= 10$ degrees on C. 14 degrees on F. $= -10$ degrees on C.

(a) What on the C. scale $= 70$ on the F. scale.

(b) What on the C. scale $= 4$ degrees *below* zero on the F. scale?

(c) What on the F. scale $= 20$ degrees on the C. scale?

O. If $a=3$ and $b=2$, what does a^2-ab equal?

P. If $x-2a+b=2x+2b-4a$, what does x equal?

Q. If $\frac{4}{x+2} + \frac{7}{x+3} = \frac{37}{x^2+5x+6}$, what does x equal?

R. Let l stand for the safe load that can be hoisted by a hemp rope. Let c stand for the circumference of a rope. If $l=100c^2$ for any hemp rope, how many pounds are a safe load for a hemp rope $2\frac{1}{4}$ inches in circumference?

S. If $a=6$ and $b=1$, what does $2ab-ab^2$ equal?

T. Find the average midnight temperature for the week in which the daily midnight temperatures were 15, 3, 0, -7 , -9 , 6, and 17 degrees.

U. If $\frac{x}{a+b} = a-b$, what does x equal?

V. How much water must be added to a pint of "alcohol, 95% pure," to make a solution of "alcohol, 40% pure"?

W. Given that $2x - 3$ is less than $x + 5$ and that $11 + 2x$ is less than $3x + 5$, to find the limits (*i. e.*, the values) between which x lies.

(It is understood that the pupil has not had any special training in inequalities or limits. This problem is, so to speak, an original exercise.)

X. At what time between 6 and 6:30 o'clock are the hands of a watch at right angles to each other?

Y. If $x = \frac{a+b}{2}$ what does $\left(\frac{x-a}{x-b}\right)^3 - \frac{x-2a+b}{x+a-2b}$ equal?

INSTRUCTIONS FOR THE EXPERIMENT.

Cut up the sheet of problems so as to have twenty-five strips, each strip with one problem. Examine the problems with a view to ranking them in order of difficulty for pupils of age from 14 years 0 months to 15 years 11 months, who have had twenty weeks' work in algebra, five periods a week, or its equivalent. Mark the easiest one, 1; mark the next easiest one, 2; mark the next easiest, 3; continue, the hardest problem being marked 25. If two or more of the problems seem to be of equal difficulty, choose one of them at random for the easiest of them, and so on. That is, suppose that you have already picked out as to difficulty your Nos. 1, 2, 3, 4, and 5, and that you then find three problems that seem equally difficult. These are to fill ranks 6, 7 and 8. Shuffle them, take one for 6, another for 7, and another for 8, at random.

In ranking the problems let "more difficult" mean, for any example, "likely to be solved correctly *in thirty minutes* by a smaller percentage of the pupils." That is, assume that thirty minutes are allowed for each problem and try to prophesy the ranking of the problems by the number out of ten thousand pupils (of the sort stated above) who would fail with each problem. If you prophesy that 2,000 pupils will fail to get Problem *A* and 1,900 will fail on Problem *B*, then call *A* more difficult than *B*.

The amounts of difficulty ascribed to these problems by the consensus of the two hundred teachers can be inferred from the percentage of them judging each example to be harder than each other example. If, that is, eighty per cent. of them judged *K* to be harder than *D*, while 90 per cent. of them judged *A* to be harder than *D*, and 99 or 100 per cent. of them judged *O* to be harder than *D*, we should (letting *d*, *k*, *a*, and *o* equal, respectively, the amounts of "difficulty" possessed by *D*, *K*, *A* and *O*

in the minds of the two hundred teachers) be confident that $o > a$, that $a > k$, and that $k > d$. If, further, eighty per cent. of them judged A to be harder than K , we may be confident that $a-k$ is approximately equal to $k-d$, *since the two differences are equally often judged correctly by the two hundred judges, there being a minority above zero.*

Counting up the 200 decisions in the case of each problem's comparison with every other, we have a table of which Table II. shows the first three lines in part as a sample. This table reads: 83 per cent. of the teachers judged that $k > d$; 93 per cent. of them judged that $a > d$; $94\frac{1}{2}$ per cent. of them judged that $l > d$; 100 per cent. of them judged that $o > d$; 82 per cent. of them judged that $a > k$; $88\frac{1}{2}$ per cent. of them judged that $l > k$; 97 per cent. of them judged that $o > k$; etc.

TABLE II.

THE FREQUENCY OF JUDGMENTS THAT " K IS MORE DIFFICULT THAN D ," THAT " A IS MORE DIFFICULT THAN D ," THAT " L IS MORE DIFFICULT THAN D ," ETC., ETC. In percentages.

	K	A	L	O	S	T	P	U	G	B	C	Q	etc.
D	83	93	$94\frac{1}{2}$	100	100	etc.							
K		82	$88\frac{1}{2}$	97	97	97	etc.						
A			67	$64\frac{1}{2}$	71	$83\frac{1}{2}$	99	etc.					
L				$59\frac{1}{2}$	64	$72\frac{1}{2}$	$99\frac{1}{2}$	93	etc.				
O													
S													
etc.													

It is possible to use knowledge of the relation that holds good between (1) the percentage of judges judging a certain difference correctly and (2) the amount of the difference, so as to infer the latter from the former. I will not, however, enter into either the principles or the technique of doing so for the case in hand, since a much simpler method of defining the amounts of difficulty of a series of these problems is available.

Let a , b , c , etc., equal respectively the "difficulty" of A , "difficulty" of B , etc., in the opinions of the two hundred teachers, as heretofore. Then the table from which Table II. is an excerpt would show the following:

83 % of the judges judged that $k > d$
 82 % of the judges judged that $a > k$
 83½% of the judges judged that $t > a$
 81½% of the judges judged that $h > t$
 82½% of the judges judged that $e > h$
 81 % of the judges judged that $i > e$

Hence, approximately, $k-d=a-k$; $a-k=t-a$; $t-a=h-t$; $h-t=e-h$; and $e-h=i-e$.

Call the amount of difference in "difficulty" necessary in order that 82¼ per cent. of the judges should rate the harder of two problems as harder, *dif.*

Then, approximately,

the difficulty of D in the minds of the 200 judges is d
 the difficulty of K in the minds of the 200 judges is $d + 1 \text{ dif.}$
 the difficulty of A in the minds of the 200 judges is $d + 2 \text{ dif.}$
 the difficulty of T in the minds of the 200 judges is $d + 3 \text{ dif.}$
 the difficulty of H in the minds of the 200 judges is $d + 4 \text{ dif.}$
 the difficulty of E in the minds of the 200 judges is $d + 5 \text{ dif.}$
 the difficulty of I in the minds of the 200 judges is $d + 6 \text{ dif.}$

The difficulty of D , the difficulty of K , the difficulty of A , and so on, can now be stated as we state the weights of certain objects, or the amounts of wealth possessed by certain men, *if we can estimate the amount of difficulty of D itself*—that is, the difference between the difficulty of D and zero, or "just not any," difficulty.

A problem in algebra that approached zero difficulty might be variously defined, but perhaps the most useful definition would be: "*A problem that is truly algebraic, not simply a problem for quantitative thinking in general, and certainly not simply a problem for thought in general; and that has some difficulty, but less than any other such problem.*" For example, the problem, "Call one dime d , call two dimes $2d$, call three dimes $3d$. How many cents are there in $4d$?" is truly algebraic and would, by many competent judges, be rated as easier than D . The problem, "Call one dime d , call two dimes $2d$, call three dimes $3d$. How many cents are there in $3d$?" is also truly algebraic and is perhaps still easier. Somewhere amongst the algebraic problems that can be devised that are easier than D will be found

one (call it α) nearly at the point of "just not any" difficulty. By "nearly" is of course meant that from the true zero of difficulty to the difficulty of α is a small fraction of the difference between the true zero and, say, the difficulty of E or I .

The difference between "zero difficulty" as an algebra problem and the difficulty of D has not been determined. It is almost surely, in the opinion on the two hundred, less than 2 *dif.* and probably a little, if at all, over 1 *dif.* Call it 1 *dif.* That is, assume that zero or "just not any" difficulty as a problem in algebra is represented by a problem as much easier than D as D is easier than K .

Then

the difficulty of $D = 1$ *dif.*
 the difficulty of $K = 2$ *dif.*
 the difficulty of $A = 3$ *dif.*
 the difficulty of $T = 4$ *dif.*
 the difficulty of $H = 5$ *dif.*
 the difficulty of $E = 6$ *dif.*
 the difficulty of $I = 7$ *dif.*

And T is twice as difficult as K in the sense of being twice as far from zero on a scale for "difficulty" as defined by the consensus; E is one and a half times as difficult as T , twice as difficult as A ; etc., etc.

By a fairly permissible hypothesis about the nature of the individual differences of these 200 teachers' powers of judging these differences in difficulty, the constant which I have here called "1 *dif.*" equals 1.371 times the amount of added difficulty which would cause 75 per cent. of these teachers to rate the harder of two problems as harder.* Since in other work with educational scales, this "*P.E.*" or "75%-difference" has been taken as a unit, we may well restate the scale above as

* An account of the nature of this hypothesis and the consequent derivation of the 1.371 *P.E.* is beyond the scope of this paper. Its essential feature is the assumption that the factors which make the judgments of two hundred teachers of mathematics concerning a problem's difficulty vary one from another can be arranged in n groups, these groups being approximately equal in magnitude of influence, and approximately uncorrelated in action, and that n is fairly large (say, 15 or more).

the difficulty of $D = 1.37$ P.E.
 the difficulty of $K = 2.74$ P.E.
 the difficulty of $A = 4.11$ P.E.
 the difficulty of $T = 5.48$ P.E.
 the difficulty of $H = 6.85$ P.E.
 the difficulty of $E = 8.22$ P.E.
 the difficulty of $I = 9.59$ P.E.

73½ per cent. of the two hundred judges judged that $v > i$; and 78½ per cent. of them, that $w > v$. For reasons which I shall here omit, we may infer that, approximately,

$$v - i = .71 \text{ dif. or } .98 \text{ P.E.}$$

$$w - v = .85 \text{ dif. or } 1.17 \text{ P.E.}$$

and

$$v = 7.7 \text{ dif. or } 10.57 \text{ P.E.}; w = 8.55 \text{ dif. or } 11.74 \text{ P.E.}$$

So our entire series of standards, by the consensus of the two hundred judges, may be taken roughly as

d , the difficulty of D , = 1.4 P.E. or 1 dif.
 k , the difficulty of K , = 2.7 P.E. or 2 dif.
 a , the difficulty of A , = 4.1 P.E. or 3 dif.
 t , the difficulty of T , = 5.5 P.E. or 4 dif.
 h , the difficulty of H , = 6.9 P.E. or 5 dif.
 e , the difficulty of E , = 8.2 P.E. or 6 dif.
 i , the difficulty of I , = 9.6 P.E. or 7 dif.
 v , the difficulty of V , = 10.6 P.E. or 7.7 dif.
 w , the difficulty of W , = 11.7 P.E. or 8.5 dif.

These results are sound as an expression of the approximate amounts of "difficulty" of the problem in question in the opinion of the two hundred teachers in question. And such a consensus as treated gives estimates of these amounts of "difficulty" far superior to the opinion of a single teacher. Until a better scale is devised this scale may be freely used as a foot-rule to define and measure the difficulty of examples in algebra for pupils who have studied the subject for the stated length of time.

By objective tests of pupils a better scale can be devised. Its validity will depend upon the validity of the assumptions made concerning the distribution of ability to solve problems in algebra in the pupils in question. The general nature of the argument in such a case has been fully illustrated by Dr. Buckingham in the case of a scale for the spelling-difficulty of words.

It will have been understood, I trust, that the extent to which "difficulty" as judged by these two hundred teachers is identical with "difficulty" as judged by a consensus of, say, English or French, or Canadian, or Californian, teachers is a matter to be determined by experiment; and, similarly, that the extent to which any ten thousand children tested represent any other group of children is a matter to be determined by experiment. "Difficulty" is, of course, not one thing, but many, varying for the same problem with the amount of training and its nature. The scale, as given, has just the extent of application that this particular consensus warrants. Probably no one realizes the limitations of these measures of the "difficulty" of D , K , A , etc., so well as the author, who can easily add three to every one of the criticisms of such scales that he has received. Yet I unhesitatingly assert that, at the present time, no measures of the "difficulty" of D , K , A , etc., for pupils after 20 weeks' work in algebra exist that have one half the probability of being true that the measures stated here have.

INDIVIDUAL DIFFERENCES AMONGST TEACHERS IN RESPECT TO
AGREEMENT WITH THE CONSENSUS IN RATING 24 PROBLEMS,
 G BEING OMITTED FROM CONSIDERATION.

The order assigned to the problems by any one teacher must, Table I being the facts, as a rule differ from the order assigned by the consensus. I have computed the facts for a hundred of the teachers. In computing them I have taken the differences (regardless of signs) between the teachers's ratings and the corresponding ratings by the consensus, and found the sum for each teacher. In this sum G does not figure, since the variation was so largely due to varying opinions concerning whether the pupils would have been taught the meaning of the $\sqrt{}$ sign. Differences for T and P from *either* 7 or 8 (whichever was nearest) were used; differences for I , X and Y from 18 *or* 19 *or* 20 (whichever was nearest) were used.

The resulting sum of differences from the consensus ranges from 16 to 97, the latter result being rather far on toward the result that would be got from a random shuffling of the problems! The details are shown in Table III.

TABLE III.

INDIVIDUAL DIFFERENCES IN AMOUNT OF DISAGREEMENT WITH THE CONSENSUS IN THE ORDER OF DIFFICULTY OF 24 PROBLEMS.

Quantity. Sum of Differences from the Consensus of the 200 Judges.	Frequency, Percentage of Teachers.
16-23	3
24-31	10
32-39	18
40-47	29
48-55	18
56-63	14
64-71	4
72-79	2
80-87	1
88-95	0
96-103	1

These sums of differences represent (inversely) complexes of (1) ability to judge as the consensus does when it is *right*—i. e., real ability to judge the “difficulty” of problems, (2) acquaintance with conditions as to the curriculum in algebra that are nearer the average condition, (3) ability to judge as the consensus does when it is *wrong*, (4) time and care given to the ratings, and (5) accidental or “chance” variations due (5a) to the individual’s condition at the time and (5b) to the small number of problems judged.

The last (5b) is not a very important cause of the differences found amongst teachers. For the unreliability due to sampling, though fairly large, is small in comparison with the individual differences themselves. For example the two halves of the series (*A* to *K* and *L* to *Y*) gave for the five “best” and five “worst” individuals the following results:

Five Best.			Five Worst.		
Sums of Differences.		Δ	Sums of Differences.		Δ
<i>A</i> to <i>K</i> .	<i>L</i> to <i>Y</i> .		<i>A</i> to <i>K</i> .	<i>L</i> to <i>Y</i> .	
8	8	0	39	32	7
11	8	3	31	42	11
8	14	6	35	43	8
11	13	2	44	39	5
17	9	8	39	58	19

And, in general, the mean square error of any one of the entries of Table III. is, so far as the sampling of problems goes, on the average under 5.

It is my opinion that, were all causes of variation save (1) removed, the individual differences would still be so great that the sum of differences from the correct order would still show a range of over three to one.

TEACHERS COLLEGE,
COLUMBIA UNIVERSITY,
NEW YORK.